Automatic age detection in normal and pathological voice

J-A. Gómez-García , L. Moro-Velázquez , J-I. Godino-Llorente , G. Castellanos-Domínguez

Abstract

Systems that automatically detect voice pathologies are usually trained with recordings belonging to population of all ages. However such an approach might be inadequate because of the acoustic variations in the voice caused by the natural aging process. In top of that, elder voices present some perturbations in quality similar to those related to voice disorders, which make the detection of pathologies more troublesome. With this in mind, the study of methodologies which automatically incorporate information about speakers' age, aiming at a simplification in the detection of voice disorders is of interest. In this respect, the present paper introduces an age detector trained with normal and pathological voice, constituting a first step towards the study of age-dependent pathology detectors. The proposed system employs sustained vowels of the Saarbrücken database from which two age groups are examinated: adults and elders. Mel frequency cepstral coefficients for characterization, and Gaussian mixture models for classification are utilized. In addition, fusion of vowels at score level is considered to improve detection performance. Results suggest that age might be effectively recognized using normal and pathological voices when using sustained vowels as acoustical material, opening up possibilities for the design of automatic age-dependent voice pathology detection systems.

Index Terms: age recognition, pathological voice, sustained vowels

1. Introduction

Systems that automatically detect pathologies using the speech present potential advantages with respect to traditional detection and evaluation procedures by providing inexpensive, non invasive and objective means of assessing the clinical state of patients [1]. Nonetheless, the usage of speech signals poses a difficulty due to the intrinsic variability of the voice which compromises the potential performance of automatic detection systems. One of such variability sources is in the age of speakers, which has been identified as a confounding factor in the analysis of pathological voices with subsequent consequences for the assessment and treatment of voice disorders [2]. However, distinguishing between pathology and age-related changes can be difficult [3]. Indeed, some parameters typically associated with voice pathologies (such as jitter and shimmer) have proven to be highly correlated with elder voices [4]. Moreover, aging voice is described perceptually by tremor, hoarseness, voice breaks or breathiness [5], factors which might correlate to some degree with pathological conditions. All these facts indicate that aging related effects should be considered, specially when dealing with elder speakers, to avoid confusions in pathological voice detection systems.

In this respect, the voices of older adults are affected by hormonal changes and by natural deterioration processes affecting the speech production mechanisms. At hormonal level, females suffer a dramatic fall of estrogens during menopause, while males experience a gradual decline of testosterone concentrations during aging [6]. At physiological level and triggered by hormonal changes, the deterioration processes include structural alterations in the organs of speech production and in the neural mechanisms of control of these organs [2]. On one hand, the structural changes are mainly reflected in the respiratory system, the larynx and vocal folds. Regarding the respiratory system, changes include a decreased lung capacity, stiffening of the thorax and weakening of respiratory muscles. In the larynx, atrophy occurs in all intrinsic laryngeal muscles as well as an ossification of cartilages (starting in the fourth decade for females and the third for males) and calcification. In addition, it has been observed a slight lowering of the larynx in the neck that increases the length of the vocal tract [7]. Regarding the vocal folds, changes comprise a general degeneration and structural atrophy, including a shortening of males' vocal folds and alterations in females' vocal-fold closure configuration during normal phonation [8]. On the other hand, neuromotor effects are observed in both the peripheral and the central nervous system. These may affect speech rate, coordination of articulators, breath support and regulation of fundamental frequency ($F0$) [7]. Acoustically, these aging effects are explained by differences in pitch, loudness, and quality, compared to younger populations [5]. Numerous studies indicate for instance that pitch increases slightly in males and decreases more prominently in females as a function of aging [9]. More consistent measures of vocal activity and aging involve $F0$ variability which increases with age for both sexes and indicate a reduced laryngeal control [2]. Formant frequencies in vowels have been reported to lower with female and male owing to increased vocal tract length [7]. Besides, Jitter and shimmer have been found to be significant in studies with older males and females [2].

The aforementioned anatomical and acoustical cues suggest that age might be automatically recognized by means of the speech. In fact, attempts have been reported in literature. A system using cepstral and $F0$ features for characterization and a hidden Markov model for classification, is presented in [10]. By using continuous speech, a 70% classification accuracy is obtained in differentiating between young, adult and elder speakers. In [4], jitter and shimmer are employed as features along with 6 classifiers. By using artificial neural networks and continuous speech, the accuracy in classifying between elder and non elder speakers is 96.57%. The authors also incorporate sex information in their system by means of an automatic sex detector and a Bayesian network. An automatic speaker age and sex identification system using acoustic and prosodic features is presented in [11]. Seven different subsystems based on support vector machines, Gaussian mixture models (GMM) and

supervectors are tested out. At the end, a fusion at score level of the different subsystem performed up to 52% in differentiating between 4 age groups. In [12], a system based on $i-$vectors and least squares support vector regression is employed to estimate the age of speakers. Using continuous speech the proposed method provided relative improvements of 5% mean absolute error compared to their best baseline.

Having those precedents, the present paper aims at extending previous attempts of automatically recognizing speaker's age but using normal and pathological speech, towards the design of an automatic age-dependent voice pathology recognition system. A simple methodology based on Mel frequency cepstral coefficients (MFCC) and GMM classifiers is considered, since the objective is not to find the system that provides the best performance in automatic age recognition, but to test out if it is feasible to extract this age information directly from normal and pathological voice recordings. Experiments are performed using registers of the sustained phonation of vowels $/a/$, $/i/$ and $/u/$. This implies the design three subsystems, one for each vowel, that are latter fused at score level by means of logistic regression. Finally, and to verify the influence of sex in the age detection task, a sex-dependent age detector is considered.

The structure of the paper is as follows: Section 2 introduces the theoretical background. Section 3 introduces the experiments, the database and the methodology. Section 4 presents the results, while section 5 introduces the discussions, conclusions and future work.

2. Theoretical Background

2.1. Logistic regression fusion

Having a feature vector $\boldsymbol{x}$ of dimension d, belonging to a particular class j, a GMM is defined as a finite mixture of G multivariate Gaussian components of the form:

$$p(\boldsymbol{x}|\Theta^j) = \sum_{r=1}^{G} \boldsymbol{\lambda}_r^j \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_r^j, \boldsymbol{\Sigma}_r^j)$$

where $\boldsymbol{\lambda}_r^j$ are scalar mixture weights, $\mathcal{N}(\cdot)$ are Gaussian density functions with mean $\boldsymbol{\mu}_r^j$ of dimension d and covariances $\boldsymbol{\Sigma}_r^j$ of dimension $d \times d$, and $\Theta^j = \{\boldsymbol{\lambda}_r^j, \boldsymbol{\mu}_r^j, \boldsymbol{\Sigma}_r^j\} \mid_{r=1}^{G}$ comprises the above mentioned set of parameters. Θ_i can be estimated using the *expectation-maximization* algorithm in a *maximum likelihood* maximization scheme. Since the design of an age detector is a two-class problem (elder vs. adult), two models are required, one representing the target class, $p(\boldsymbol{x}|\Theta^c)$, and the other the complementary class, $p(\boldsymbol{x}|\Theta^{\bar{c}})$. Now, it is possible to derive a probabilistic output of the membership of some new data $\boldsymbol{y}$ to each one of the models, $p(\boldsymbol{y}|\Theta^j)$. This permits to derive a log-likelihood decision function, $\Lambda(\cdot)$, for discriminating if the new data belongs to the target class or to the complementary. This decision function is as follows:

$$\Lambda(\boldsymbol{y}) = \log \frac{p(\boldsymbol{y}|\Theta^c)}{p(\boldsymbol{y}|\Theta^{\bar{c}})}$$

Thus, the decision is taken establishing a threshold over $\Lambda(\cdot)$ values obtained for each vector to be classified.

In order to incorporate information of different recognizers into a single system, a fusion procedure is employed as in [13, 14]. Thus, let there be K input recognizers, where the $k-th$ recognizer outputs its own log-likelihood-vector $\Lambda_k(\boldsymbol{y}_t)$ for a given trial $\boldsymbol{y}_t$ of a set of T trials. A new recognizer, Λ^*, is formed as follows:

$$\Lambda^*(\boldsymbol{y}_t) = \sum_{k=1}^{K} \alpha_k \Lambda_k(\boldsymbol{y}_t) + \boldsymbol{\beta}$$

where α_k are scalar weights, and $\boldsymbol{\beta}$ is an vector performing an affine calibration transformation on the fusion output.

The parameters $\gamma = [\alpha_1, \cdots, \alpha_K, \boldsymbol{\beta}]$ are optimized with logistic regression, maximizing the objective function C_{llr} on an evaluation database. The optimization function for a two-classes problem is as follows:

$$C_{llr} = \frac{0.5}{|c|} \sum_{t \in c} \log_2 \left(1 + e^{-\Lambda(\boldsymbol{y}_t)}\right) + \frac{0.5}{|\bar{c}|} \sum_{t \in \bar{c}} \log_2 \left(1 + e^{\Lambda(\boldsymbol{y}_t)}\right)$$

3. Experimental Set-Up

3.1. Database

The Saarbrücken Voice Database [15, 16] holds a collection of voice signals from more than 2000 normal and pathological German speakers. It contains recordings of the sustained phonation of vowels $/i/$, $/a/$, and $/u/$ produced at normal, high and low pitch, as well as with rising-falling pitch. Voice recordings were obtained using the Computerized Speech Lab (CSL) station $4300B$, using a sampling frequency of $50\,kHz$ and 16-bits of resolution. A subset of the database was segmented by a speech therapist, removing those recordings with a low dynamic range or interference. Additionally, speakers aged less than 16 are discarded.

For this work purposes a subset of the database, consisting of the vowels $/i/$, $/a/$, and $/u/$ pronounced at normal pitch is employed. In the age detection task two age groups are examined: adults and elders. The adult group includes normal and pathological speakers younger than 59 if male, and younger than 49 if female, while the elder group comprises the remaining speakers. This distinction among sexes is because of the differences between the aging process of females and males. In this respect, the most significant voice changes occur abruptly in females after menopause, around age 50 [17]. In males testosterone levels decrease slowly with age and thus changes are not that abrupt. To take into account this gradual hormonal degradation process in males, the age of the male elder group is set to 60. Table 1 summarizes the distribution by age and condition of the tested database.

Table 1: Saarbrücken database speaker's distribution

	Distribution by Sex		**Distribution by Condition**		
	Female	**Male**	**Normal**	**Path.**	**Total**
Adults	660	501	548	613	1161
Elders	252	132	20	364	384
Total	912	633	568	977	1545

3.2. Setup

Two subsystems are designed in this paper to classify between adult and elder speakers. On one hand, a *sex-independent age recognizer* trained on data of both female and male speakers is presented. On the other, a *sex-dependent age recognizer* is introduced to study the influence of including information about the sex of the speaker in the age recognition task. Hence, a male and female age recognizer are designed by using voices belonging to each one of sexes.

Since the differences between both systems are on the manner on which the input database is employed, the same age recognition methodology is followed. Thus, Fig. 1 presents an outline of the age recognizer developed in this paper. Each one of its stages is detailed next.

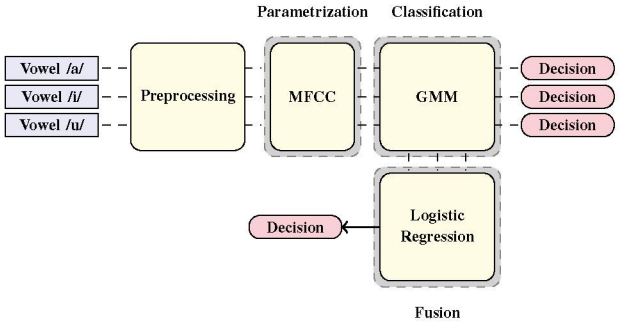


Figure 1: Outline of the automatic gender- detection system utilized in this paper

3.2.1. *Preprocessing*

All speech signals are down-sampled to $25\,kHz$, and $[-1, 1]$ normalized. Short time analysis is carried out using 50% overlapped Hamming windows $40\,ms$ long [18].

3.2.2. *Characterization*

MFCC features are employed for characterization, varying the number of coefficients by the following set: $\{4, 8, 12, 14, 16, 18, 20, 22\}$. Experiments are performed on each one of the single vowels $/a/$, $/i/$ and $/u/$.

3.2.3. *Classification*

Due to the age imbalance in the database, accuracy might not be the best measure of performance because results could be biased towards the majority class. To avoid such a problem, the area under the ROC curve (AUC) is preferred as a measure of performance assessment. [19]. AUC quantifies the overall quality of a classifier by calculating the fraction of the total area that falls under a Receiver Operation Characteristic curve. Larger AUC values indicate generally better classifier performance [19]. Along with AUC, detection error trade-off curves (DET) are also employed for assessment. The classifier accuracy, α, is also calculated within a given confidence interval q. So, assuming a 95% value, the q range is estimated as $q = \pm 1.96\sqrt{\alpha(1-\alpha)/N}$, where N is the total number of classified patterns. Additionally, specificity (s_p) and sensitivity (s_e) are computed. The measures α, s_p and s_e are determined at the Equal error rate (EER), after having selected the parameters (MFCC coefficients and number of Gaussians) which performed the best as by the AUC.

The performance evaluation includes an 11-fold cross-validation strategy, using a simple GMM for classification. The number of Gaussians is varied with the values: $\{4, 8, 16, 24, 32, 64, 100, 128\}$.

3.2.4. *Fusion*

Logistic regression is employed for fusing the scores belonging to each one of the single wovels and to obtain a single estimation of age detection. The fusion has been performed using the BOSARIS toolkit [13].

4. Results

The DET curve for the *sex-independent age detector* is presented in Figure 2, whereas Table 2 summarizes the obtained results in terms of α, s_e and s_p at the EER threshold.

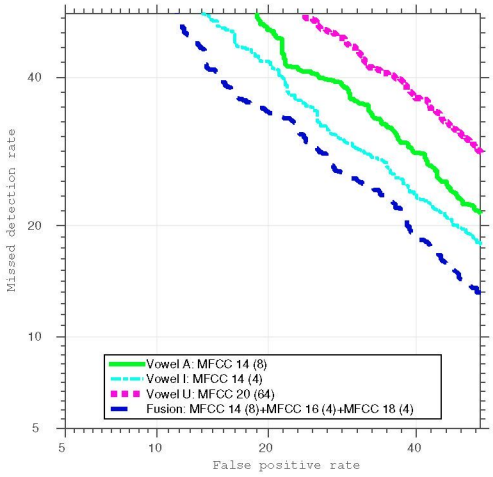


Figure 2: DET curve for the sex-independent age detector. The legend shows in parentheses the number of gaussians used for each classifier and the number of MFCC coefficients that reported the best results for each one of the tested vowels.

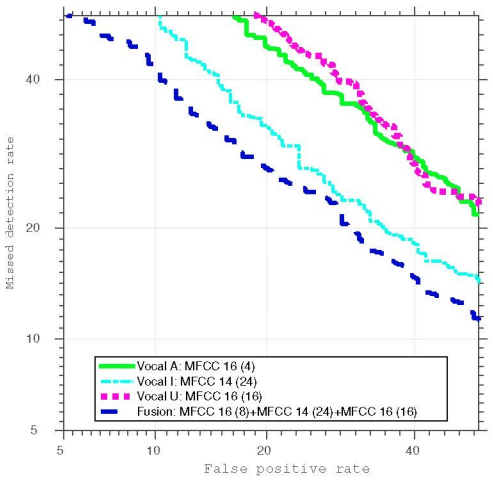
Table 2: Performance of the sex-independent age detector using the Saarbrüecken database

Vowel	***AUC***	$\alpha \pm q$	s_e	s_p
Vowel */a/*	0,75	67,51 ± 2,34	0,68	0,66
Vowel */i/*	0,77	70,10 ± 2,28	0,71	0,68
Vowel */u/*	0,70	65,50 ± 2,37	0,69	0,58
Fusion	0,80	72,49 ± 2,23	0,74	0,69

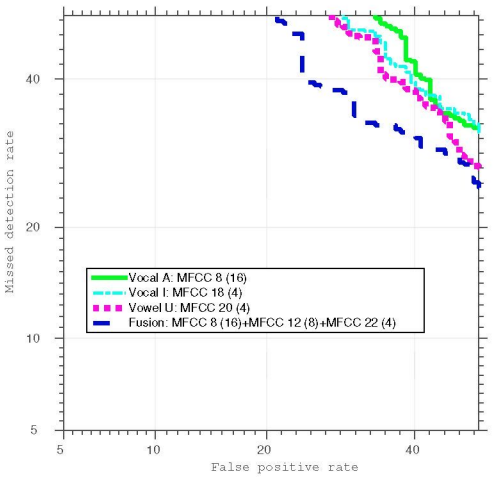
The DET curves for the *sex-dependent age detector* are presented in Figure 3, for both male and female models. Table 3 summarizes the obtained results in terms of α, s_e and s_p at the EER threshold. Results are given for male, female and mixed sexes, where the latter refers to combining male's and female's performance measures.

Table 3: Performance of the sex-dependent age detector using the Saarbrüecken database

	Vowel	***AUC***	$\alpha \pm q$	s_e	s_p
Female	Vowel */a/*	0,76	68,64 ± 3,01	0,70	0,64
	Vowel */i/*	0,81	76,32 ± 2,76	0,81	0,65
	Vowel */u/*	0,73	67,76 ± 3,03	0,71	0,59
	Fusion	0,84	77,52 ± 2,71	0,83	0,64
Male	Vowel */a/*	0,69	68,09 ± 3,63	0,73	0,51
	Vowel */i/*	0,65	61,77 ± 3,79	0,65	0,48
	Vowel */u/*	0,69	66,98 ± 3,66	0,69	0,58
	Fusion	0,73	69,51 ± 3,59	0,75	0,48
Mixed	Fusion	–	74,24 ± 2,18	0,80	0,58



(a) Female age detector



(b) Male age detector

Figure 3: DET curves for the sex-dependent age detector. (a) Female age detector. (b) Male age detector.

5. Discussion and Conclusions

The present work studies the performance of an automatic age recognizer using normal and pathological voices. The proposed methodology is based in MFCC features and GMM classifiers, and is applied to the sustained phonation of /a/, /i/ and /u/ vowels. To improve performance even further, a fusion at score level of the information provided by every single vowel is considered. One sex-independent age detector and a sex-dependent age detector are designed following the above mentioned methodology. The use of both detectors aims at investigating the influence of including or not a-priori information about the sex of the speaker in the age recognition task.

Results suggest that the proposed methodology is proficient in automatically detecting age on normal and pathological voice. In terms of AUC, the best result for the sex-independent age detector is 0.80, compared to 0.84 for the female case and 0.73 of the male case in the sex-dependent scenario. Results also indicate that a fusion of the information derived in each one of the vowels provides an improvement in performance on both cases, as might be observed in the DET curves of Figures 2 and 3. No single vowel can be regarded as the most informative, since in the different experiments the best performance favored the usage of one or another.

When doing a direct comparison between the sex-independent and the sex-dependent age detector (in the mixed scenario), a slight accuracy improvement is observed in favor of the sex-dependent system. However, such a result is not conclusive as regarded by s_e and s_p which somehow favor the sex-independent case. It should be pointed out that AUC is the criteria employed for finding the best parameters in terms of MFCC coefficients and number of gaussians. It intends to comprise into a single number different operation points, in order to simplify comparison between different systems. Nonetheless, this also implies that there might exist individual operation points on which α, s_e and s_p might provide a better performance. In view of the results, a direct comparison between the sex-independent and the sex-dependent systems is difficult, and no clear conclusions could be made about whether including or not information regarding the sex of the speaker improve performance.

Following the discussion on the sex-dependent system, the female detector seems to outperform to the male one in terms of AUC. One possible hypothesis explaining such a behavior might be in the definition of the limits for adults and elder speakers (49 in females and 59 in males). In that regard, while the menopause causes a sudden drop in hormonal levels which in turn produce the changes in the speech mechanisms, the hormonal changes in man are gradual and encompass a larger deviation in terms of years. Therefore, the definition of a frontier for elder males is somewhat fuzzier than in females and might produce higher errors. Another important point to consider is that the voice of a speaker might not correlate chronologically with its age, in such a way that a person's voice might be functionally more similar to that of a younger or of an older speaker. In this regard, non-pathological perturbations, such as those of the psychological, physiological, neurological, behavioral or external type, alter the voice qualities and difficult the establishment of limits between adult and elder speakers. Therefore, a regression procedure as the proposed in [20] might be considered to avoid troubles in the definition of age-group boundaries.

Finally and in view of the results, the presented methodology provides an effective way of distinguishing between adult and elder speakers in normal and pathological voices. The fusion of vowels increased classification performance in both sex-dependent and sex-independent detection. Since no clear conclusions whether or not including information about the sex of the speaker influence the age recognition task, more experimentation is necessary to gain understanding in the topic. Similarly, a study considering the age estimation as a regression problem and a deeper investigation on how to define the limits between adult and elder speakers are of interest. In addition, the present study has been confined to using sustained vowels as acoustic material. However, the literature reports high accuracy rates in age detection when using continuous speech. In this context, features based on prosody (such as speech rate) might also be considered in the detection task. However, the study of methodologies using such acoustic material remain as future work.

6. Acknowledgements

This research was carried out under grants: *Ayudas para la realización del doctorado (RR01/2011)* from Universidad Politécnica de Madrid and TEC2012-38630-C04-01 from the Spanish Ministry of Education.

7. References

[1] J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, S. Aguilera-Navarro, and P. Gómez-Vilda, "An integrated tool for the diagnosis of voice disorders." *Medical engineering & physics*, vol. 28, no. 3, pp. 276–89, May 2006.

[2] J. D. Harnsberger, W. S. Brown, R. Shrivastav, and H. Rothman, "Noise and tremor in the perception of vocal aging in males," *Journal of Voice*, vol. 24, no. 5, pp. 523–530, 2010.

[3] R. D. Kent, *The MIT encyclopedia of communication disorders*. MIT Press, 2004.

[4] C. Müller, F. Wittig, and J. Baus, "Exploiting speech for recognizing elderly users to respond to their special needs." in *Eighth European Conference on Speech Communication and Technology*, 2003.

[5] C. R. Hooper and A. Cralidis, "Normal changes in the speech of older adults: you've still got what it takes; it just takes a little longer!" *Perspectives on Gerontology*, vol. 14, no. 2, pp. 47–56, 2009.

[6] M. Gugatschka, K. Kiesler, B. Obermayer-Pietsch, B. Schoekler, C. Schmid, A. Groselj-Strele, and G. Friedrich, "Sex hormones and the elderly male voice." *Journal of voice: official journal of the Voice Foundation*, vol. 24, no. 3, pp. 369–73, May 2010.

[7] S. Schötz, "Acoustic analysis of adult speaker age," *Speaker Classification I*, pp. 88–107, 2007.

[8] M. Södersten, S. Hertegård, and B. Hammarberg, "Glottal closure, transglottal airflow, and voice quality in healthy middle-aged women." *Journal of Voice*, vol. 9, no. 2, pp. 182–197, 1995.

[9] S. A. Xue and D. Deliyski, "Effects of Aging on Selected Acoustic Voice Parameters: Preliminary Normative Data and Educational Implications," *Educational Gerontology*, vol. 27, pp. 159–168, 2001.

[10] I. Shafran, M. Riley, and M. Mohri, "Voice signatures," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 31–36.

[11] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[12] M. H. Bahari, M. McLaren, H. Van hamme, and D. A. van Leeuwen, "Speaker age estimation using i-vectors," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 99–108, 2014.

[13] N. Brümmer and E. de Villiers, "The bosaris toolkit user guide: Theory, algorithms and code for binary classifier score processing," *Documentation of BOSARIS toolkit*, 2011.

[14] N. Brümmer, L. Burget, J. H. Černocký, O. Glembek, F. Grézl, M. Karafiát, D. a. Van Leeuwen, P. Matějka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[15] "Saarbruecken voice database." [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/index.php4

[16] M. Putzer and W. J. Barry, "Instrumental dimensioning of normal and pathological phonation using acoustic measurements." *Clinical linguistics & phonetics*, vol. 22, no. 6, pp. 407–20, Jun. 2008.

[17] E. D'haeseleer, H. Depypere, S. Claeys, J. Van Borsel, and K. Van Lierde, "The menopause and the female larynx, clinical aspects and therapeutic options: A literature review," *Maturitas*, vol. 64, pp. 27–32, 2009.

[18] N. Saenz Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120–128, Apr. 2006.

[19] G. M. Weiss and F. Provost, "Learning when training data are costly: the effect of class distribution on tree induction," *Journal of Artificial Intelligence Research*, pp. 315–354, 2003.

[20] G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 1975–1985, 2011.